Atty. Docket No.: AEOMICA-1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: PENN, Sharron G. *et al.*

Serial No.: 09/774,203 Confirm. No.: 7320

Filed: January 29, 2001

For: METHODS AND APPARATUS FOR PREDICTING, CONFIRMING, AND DISPLAYING FUNCTIONAL INFORMATION DERIVED FROM GENOMIC SEQUENCE

Palo Alto, CA
August 16, 2001

Hon. Commissioner for Patents
Washington, D.C. 20231

PETITION UNDER 37 C.F.R. § 1.102(d)
and
M.P.E.P.§ 708.02 (II)

Sir:

Pursuant to 37 C.F.R. § 1.102(d) and M.P.E.P. § 708.02(II) (7th ed., rev. 1), applicants hereby petition to make special the above-identified application, which contains claims that applicants believe are actually being infringed in the United States. Applicants attach hereto in support of the instant Petition:

- a copy of Shoemaker *et al.*, *Nature* 409:922 - 927 (2001);
- a copy of Penn *et al.*, *Nature Genetics* 26:315 - 318 (2000); and

- the Declaration of Dr. Sharron G. Penn,

and file concurrently herewith

- the fee set forth in 37 C.F.R. § 1.17(h); and
- a Second Supplemental Information Disclosure Statement (with accompanying form PTO-1449 in duplicate and cited references).

Status of the Claims

The instant application was filed January 29, 2001, with original claims 1 - 20. On July 6, 2001, applicants filed a Preliminary Amendment* adding new claims 21 - 92. Claims 1 - 92 are presently pending and have not yet been acted upon.

Infringement

Applicants attach hereto as Exhibit A a copy of Shoemaker *et al.*, "Experimental annotation of the human genome using microarray technology," *Nature* 409:922 - 927 (15 February 2001) (the "Shoemaker reference"), which describes activities undertaken by Rosetta Inpharmatics, Inc., of Kirkland, Washington, USA. The undersigned attorney of record has made a rigid comparison of the activities described in the Shoemaker reference with claims 1 - 92 of the instant application; in the opinion of the undersigned, some of the claims, were they to issue,

---

* The Preliminary Amendment was styled a "***Second*** Preliminary Amendment Under 37 C.F.R. § 1.115" in order to distinguish it from a Preliminary Amendment earlier filed to direct entry of the Sequence Listing.

would unquestionably be infringed by the activities described by Shoemaker *et al.*

Prior art

Applicants file concurrently herewith a Second Supplemental Information Disclosure Statement, with accompanying PTO form 1449 (in duplicate) and cited references. Upon such filing, applicants will have made of record in the above-identified application the Shoemaker reference and applicant's own publication, Penn *et al.*, "Mining the human genome using microarrays of open reading frames," *Nature Genetics* 26:315 - 318 (November 2000) ("Penn reference") (also attached hereto as Exhibit B), which applicants deem the two references most closely related to the subject matter encompassed by the claims. Applicants further will have made of reference all of the references that are cited, in turn, by either the Shoemaker or Penn reference.

Applicants attach hereto as Exhibit C the Declaration of Dr. Sharron Penn, first-named inventor of the instant application and first-named author of the Penn reference. In the Declaration, Dr. Penn declares that she has a good knowledge of the pertinent prior art, and that the Shoemaker and the Penn reference are the two references deemed most closely related to the subject matter of the claims of the above-identified patent application.

Conclusion

Applicants respectfully submit that the above-identified patent application should be made special; grant of special status and an early and favorable action are respectfully requested.

Respectfully submitted,

Daniel M. Becker (Reg. No. 38,376)
Attorney for Applicants

c/o FISH & NEAVE
Customer No. 1473
1251 Avenue of the Americas
New York, NY 10020
650.617.4000

Attachments:

Exhibit A	Shoemaker *et al.*, "Experimental annotation of the human genome using microarray technology," *Nature* 409:922 - 927 (15 February 2001)
Exhibit B	Penn *et al.*, "Mining the human genome using microarrays of open reading frames," *Nature Genetics* 26:315 - 318 (November 2000) ("Penn reference")
Exhibit C	Declaration of Dr. Sharron Penn, with associated Exhibit 1

Enclosures:

- Fee, 37 C.F.R. § 1.17(h)
- Supplemental IDS with PTO 1449 (in duplicate) and copies of cited references

I Hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to Commissioner of Patents and Trademarks, Washington, D.C. 20231 on August 16, 2001

YOLANDA KEPNER
Name of Person Signing Certificate

Yolanda Kepner
Signature of Person Signing Certificate

08-16-01
Date of Signature

Experimental annotation of the human genome using microarray technology

D. D. Shoemaker*, E. E. Schadt*, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer & M. S. Boguski

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA
** These authors contributed equally to this work*

The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins. Such a catalogue cannot be constructed by computational annotation alone; it requires experimental validation on a genome scale. Using 'exon' and 'tiling' arrays fabricated by ink-jet oligonucleotide synthesis, we devised an experimental approach to validate and refine computational gene predictions and define full-length transcripts on the basis of co-regulated expression of their exons. These methods can provide more accurate gene numbers and allow the detection of mRNA splice variants and identification of the tissue- and disease-specific conditions under which genes are expressed. We apply our technique to chromosome 22q under 69 experimental condition pairs, and to the entire human genome under two experimental conditions. We discuss implications for more comprehensive, consistent and reliable genome annotation, more efficient, full-length complementary DNA cloning strategies and application to complex diseases.

The initial interpretation of a genome sequence rests upon conclusions derived solely from bioinformatics approaches—*ab initio* gene predictions, homology studies, motif analysis and other non-experimental methods[1–3]. The limitations and fallibility of this process have been discussed[4,5] and one group has concluded[6] that, despite more than 17 years of research effort[7], precise annotation of every gene in the human genome by computational methods alone is still a distant goal. Bioinformatics analyses of fragmentary experimental data have led to widely varying estimates of the number of human genes[8–10]. Comparative genomics approaches, particularly between human and mouse[11–13], will help to identify candidate genes and refine their structures, but cannot alone show that a gene is active. Consequently, projects to clone and catalogue 'full-length' cDNA clones from human[14] and mouse[15] have been undertaken. Although these projects may capture the complete coding sequences of many genes in time, cDNA cloning fixes a gene product at a particular time and under particular conditions, and thus cannot efficiently reveal the multiform nature of a metazoan transcriptome.

Recent work indicates that the human genome may contain fewer genes than anticipated[8,9], and that frequent alternative splicing might account for much physiological complexity[16–19]. This situation makes it essential to pursue a course that efficiently yields empirical validation of the structures of genes and simultaneously provides an accurate and complete catalogue of their expressed products (mRNA and cognate protein sequences).

We describe a high-throughput, microarray-based experimental method to validate predicted exons, group the exons into genes by co-regulated expression and define full-length mRNA transcripts. The method involves the design and fabrication of 'exon arrays' consisting of long (50–60 bases) oligonucleotide probes derived from predicted exons, followed by hybridization with fluorescently labelled cDNAs derived from specific cell lines or normal or diseased tissues. Absolute intensities (measuring cellular abundances) or intensity ratios (measuring differential expression regulation) from hybridized cDNAs are used to identify those probes that represent authentic exons under the conditions tested. In addition, the expression data can define gene boundaries, because adjacent exons that are co-regulated across many conditions are likely to be from the same transcript. For a higher-resolution view of gene structure, we use 'tiling arrays' in which overlapping oligonucleotides are designed to blanket an entire genomic region of interest. This approach can potentially reveal exons not identified by current gene prediction algorithms and provide information about alternative splicing.

We applied the exon array approach to a detailed analysis of human chromosome 22 under 69 pairs of experimental conditions. Tiling arrays were used to refine the structure of new genes discovered by exon analysis. Finally, a preliminary analysis of the entire human genome using exon arrays under two experimental conditions demonstrated the power of being able experimentally to validate hundreds of thousands of exon predictions, anticipating the prospect of analysing the entire human genome to a depth similar to that achieved on chromosome 22.

Analysis of chromosome 22q using exon arrays

Chromosome 22 was the first human chromosome to be completely sequenced and subjected to exhaustive computational annotation[2]. It has thus served as a benchmark for new computational and experimental methods of analysis[20,21]. We designed a single ink-jet array to monitor the 8,183 exons annotated[2] on chromosome 22q under diverse experimental conditions. Specifically, mRNAs from human cell lines and normal and diseased tissues (Fig. 1) were fluorescently labelled with two colours and hybridized in pairs to 69 individual chromosome 22 exon arrays (see Methods). Figure 2a shows a graphical display of error-weighted log expression ratios[22] for all 8,183 exons across 69 condition pairs. We developed a gene identification algorithm that uses intensity and ratio information to identify exons in a local neighbourhood that are strongly correlated across condition pairs, and then to extend such regions by incorporating other local exons with similar expression behaviour. The resulting 572 groups of co-regulated exons are referred to as expression-verified genes (EVGs). Figure 2b–e shows expanded views of specific regions of chromosome 22. Expression data can be used to confirm the exons and structure of a known gene (Fig. 2b), to identify potential false positive exon predictions (Fig. 2c), to merge UniGene clusters into a single gene (Fig. 2d) and to verify *ab initio* gene predictions experimentally (Fig. 2e).

For a chromosome-wide performance summary, we compared our experimentally derived EVGs to the list of 545 genes annotated

by Dunham *et al.*[2] (Table 1). These annotated genes were divided into four categories (known, related, predicted and *ab initio*) on the basis of the level of experimental support for the predictions. We identified 210 (85%) of the 247 known genes by analysing the expression data from the 69 condition pairs with our gene detection algorithm. The remaining 15% of known genes did not exhibit sufficient differential expression regulation among the conditions tested to enable ratio-based algorithms to verify them. We detected 66% of the related genes and 53% of the predicted genes using our expression regulation criteria. The most interesting result comes from the 325 *ab initio* genes that represented pure Genscan predictions. Dunham *et al.*[2] speculated that only 100 of these predicted transcripts would represent portions of 'real' genes, but we found experimental support for 185 (57%) of the genes in this category.

A few of the EVGs that we detected contained more than one gene. This occurred when adjacent genes were co-regulated across the 69 experimental conditions tested. In most cases, this situation can be addressed by testing additional conditions or by using additional bioinformatics techniques (for example, open reading frame (ORF) analysis, identification of internal polyadenylation sites, and supporting expressed sequence tag (EST) and protein sequence data). In a few cases, a single gene was represented by more than one EVG, indicating possible alternative splicing. Other algorithms are being developed to address this issue.

Applications of tiling arrays

Exon-based gene validation arrays can be limited by the fact that gene prediction programs perform best on 'internal' exons and not very well on initial and terminal exons, or exons that correspond to the 5′ and 3′ untranslated regions (UTRs) of mRNAs[6]. Oligonucleotide tiling arrays of overlapping probes (Fig. 3) can effectively address this challenge because they are constructed without any *a priori* knowledge of the possible exon content of a genomic sequence. We designed tiling arrays covering both strands of various genomic regions on chromosome 22 defined by EVGs where the underlying gene structure was thought to be incomplete.

Figure 3 shows how the tiling approach was used to refine the structure of the novel testis transcript described in Fig. 2e. We fabricated an ink-jet array that contained 60-mer probes spaced in 10-base-pair (bp) intervals across both strands of the 113-kilobase (kb) bacterial artificial chromosome (BAC) clone containing the EVG of interest. The array was hybridized with fluorescently labelled testis mRNA and the resulting probe intensities were analysed to determine the approximate locations of the exons within this region. For each exon, the hybridization data effectively reduced the search for the intron–exon boundaries to regions of around 20–30 bp. The exact splice junctions can generally be identified within these narrow windows by using common rules (for example, GT-AG consensus sequence and ORF analysis). For the gene shown in Fig. 3, only four of the six exons were correctly predicted by Genscan. Our results extend the 3′ UTR by 450 bp and one of the internal coding exons by 102 bases (34 amino acids). These results were confirmed by polymerase chain reaction with reverse transcriptase (RT–PCR) and sequencing (data not shown). The mRNA (GenBank accession no. AF324466) derived from this validated and corrected gene is 1,312 nucleotides long, including a 649-base 3′ UTR with a polyadenylation signal at base 1,293. It encodes a 217-residue protein and a BLASTP search revealed only one significant match (*E*-value $\sim 10^{-15}$) to a predicted gene product, CG5280 from the *Drosophila* genome project[23].

Human genome scan using exon arrays

To show that the approach described above can scale to survey the entire human genome, we used the 15 June 2000 version of the Ensembl human genome annotation data set (http://www.ensembl.org/)[24] to make 50 arrays containing 1,090,408 oligonucleotide probes representing 442,785 exons predicted by Genscan[25]. Fluorescently labelled cDNAs from a human lymphoma cell line and a colorectal carcinoma cell line were hybridized to the arrays. Analysis of fluorescence intensities from this single pair of experimental conditions provided experimental evidence for 58% of the 78,486 Ensembl confirmed exons. We detected 34% of the 364,299 predicted exons that did not meet the Ensembl 'confirmed' criteria. The false positive rate for this analysis was estimated to be around 5%, from an analysis of a set of negative control probes included on the arrays. A summary of the exons validated by this genome survey is given (see Methods) in Fig. 4 and a full listing is available as Supplementary Information or from the Rosetta website at www.rii.com.

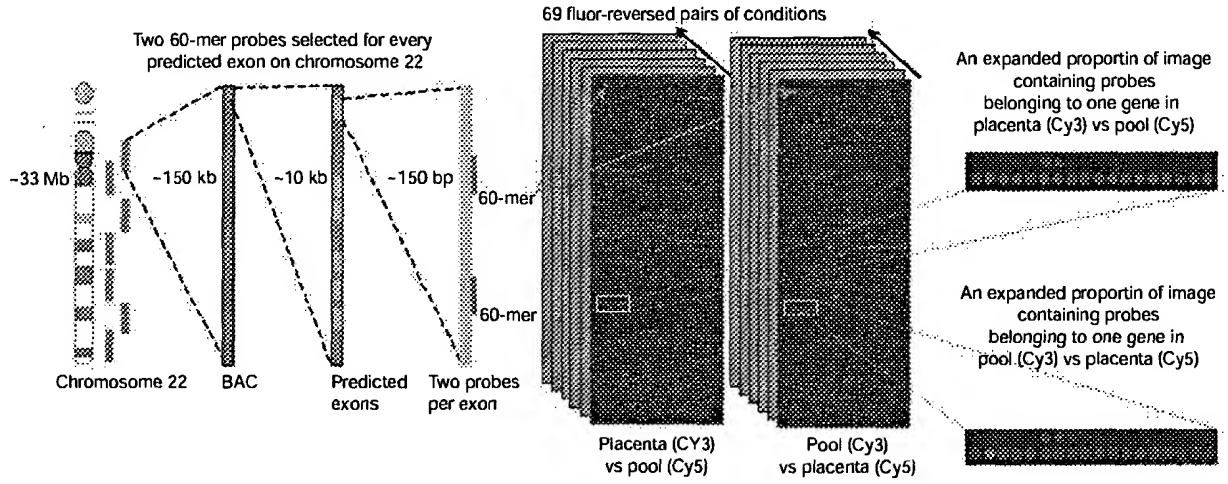

Figure 1 Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).

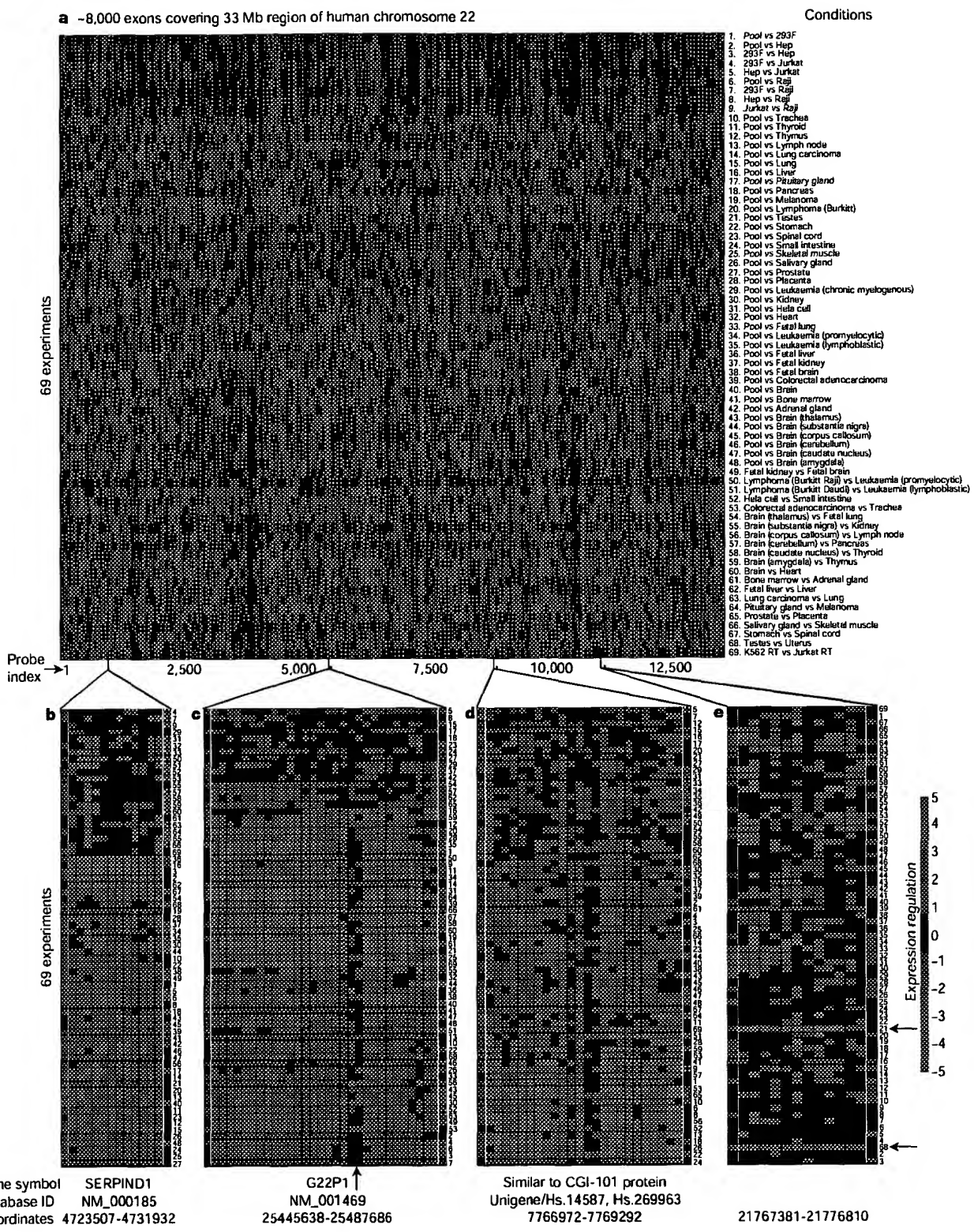

Figure 2 Using expression data from multiple conditions to validate exons and define *gene boundaries on chromosome 22.* **a**, *Pseudocolour image showing error-weighted* $\log_{10}$ expression ratios (red/green) for each of the ~8,000 exons (*x*-axis) across the 69 fluor-reversed experiments (*y*-axis). A brief description of each experiment is listed on the right side of the image; the numbers (1–69) are reference points for the Table in the Supplementary Information. The 15,511 probes representing the 8,183 predicted exons are arranged linearly across the 33 Mb of chromosome 22. **b**, Expanded region showing a known gene (SERPIND1, NM_ 000185). The experiments on the *y*-axis have been clustered to emphasize how co-regulation across diverse experiments can be used to group exons into genes. The vertical white lines indicate the boundaries predicted by our *gene finding algorithm; numbers on y-axis indicate experimental conditions.* **c**, Expanded region showing a set of co-regulated exons from another known gene (G22P1, NM_001469), illustrating the detection of potential false positives (arrow) made by the Genscan prediction program. **d**, Expanded region representing an EVG that collapses two Unigene EST clusters (HS.269963 and HS.14587) into a single transcript. **e**, Expanded region showing an EVG containing six exons that are part of a novel testis-expressed transcript (arrows, two experiments involving testis RNA samples).

Discussion

Post-genome biology and medicine will increasingly rely on complete and accurate catalogues of human genes, mRNAs and proteins. This 'parts list' is currently a patchwork of mostly hypothetical entities with varying degrees of supporting evidence. Computational techniques for sequence annotation provide invaluable clues to gene structure and function but experimental data will be required to provide a full and satisfying picture. Our microarray-based technology represents a comprehensive and consistent approach to the simultaneous validation of gene predictions and study of the transcriptome under any number of biologically or medically interesting conditions. Our approach is applicable on a genome scale and also on the scale of defining the structure of a single, novel cDNA.

The exon-based approach is well suited to high-throughput screening of diverse cell types, growth conditions and disease states. Differential expression is an important tool for assembling exons into genes. We could detect differential expression for only 15% of the confirmed exons across the human genome with a single condition pair. Clearly, larger data sets will be essential for defining the structures of genes, detecting rarely expressed genes and addressing more complex issues such as alternative splicing. In addition, information from the exon analysis can be used to select genomic regions and samples for comprehensive tiling arrays.

Ambitious efforts to clone and sequence 'full-length' cDNAs for the human[14] and mouse[15] genomes have begun with the purpose not only of helping to validate computational gene predictions but also of providing physical reagents for functional and structural geno-

Table 1 Gene validation summary of human chromosome 22q

	Annotation from ref. 2	Expression-verified genes (EVGs)	Validation fraction
Known genes*	247	210	85%
Related genes*	150	99	66%
Predicted genes*	148	78	53%
Ab initio genes*	325	185	57%

EVG sequences were searched against current versions of dbEST and nr (www.ncbi.nlm.nih.gov) and significant matches were defined as those having an *E*-value < 10^{-20}.
*Category definitions according to Dunham *et al.*[2].

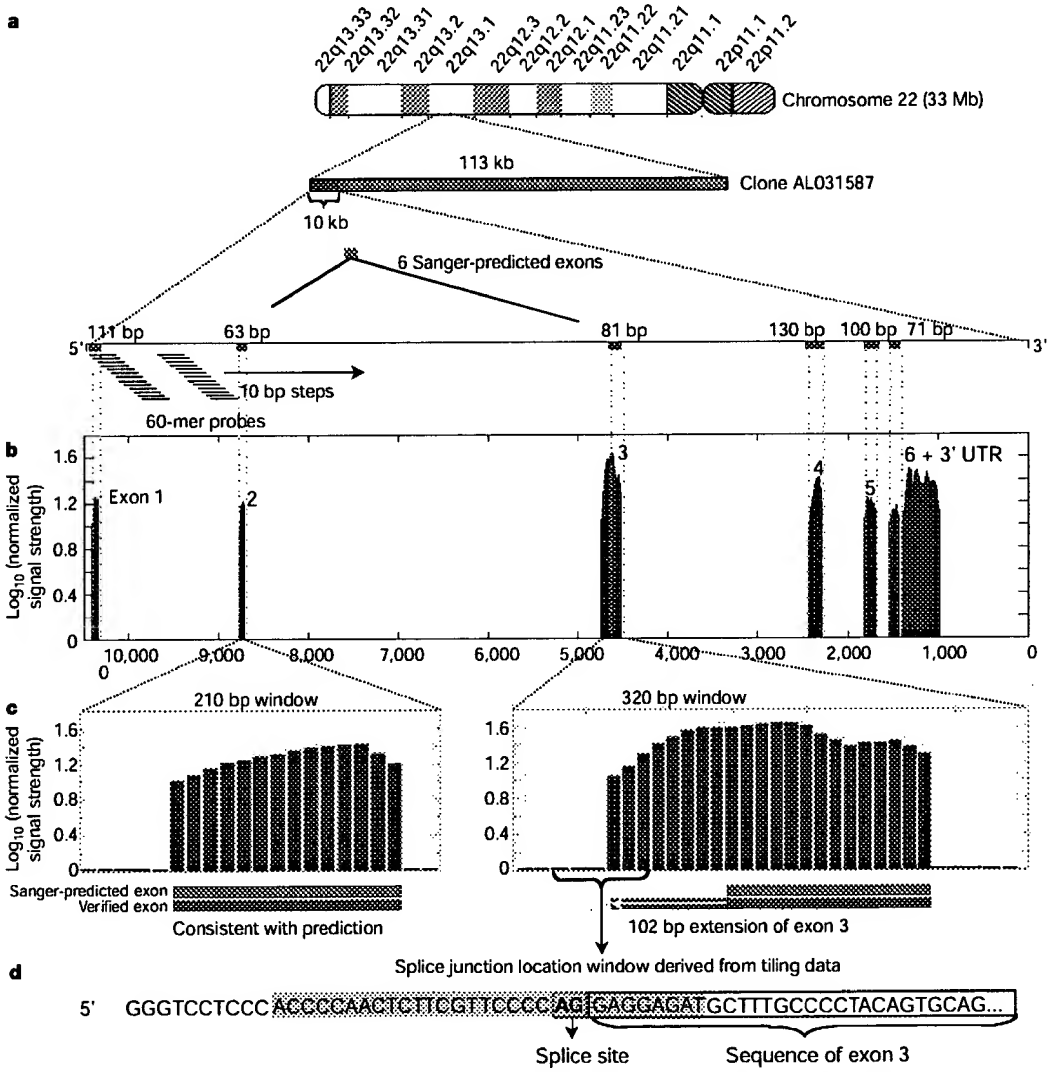

Figure 3 Characterization of a novel testis transcript using tiling arrays. **a**, An EVG discovered in the analysis of chromosome 22 (Fig. 2e) was localized to a 10-kb region at one end of the insert of BAC clone AL031587. Both strands of this 113-kb genomic interval were tiled with 60-mer probes at 10-bp steps. The tiling array was hybridized with RNA from human testis. **b**, Hybridization signals corresponding to tiling probes from this region were filtered and plotted as $\log_{10}$ values of the normalized signal strengths. Of the six Genscan predicted exons in this region, two (exons 3 and 6) were at variance with the hybridization data. **c**, Detailed views of tiling data showing one correctly predicted exon and one incorrectly predicted exon. **d**, Typically, tiling data narrow the search window for an intron/exon boundary to 20–30-bp. The exact splice junction is then identified using consensus sequences (GT-AG rule) and ORF information. The exact splice junction can also be determined by sequencing RT–PCR products.

mics. The comprehensive set of EVGs generated by our approach will accelerate these efforts by allowing a more directed cloning strategy. We also expect that hybridization data defining EVGs will be useful in 'training' the next generation of gene prediction algorithms, in much the same manner that sequence similarity data enhances *ab initio* predictions in the current state-of-the-art programs. In this way, the maximum value can be realized from the intersection of computational and high-throughput experimental biology.

Our experimental method of annotating the human genome could be rapidly reiterated for updated sequence information from the Human Genome Project, and could easily be extended to the genomes of other organisms. Generating exon and tiling arrays requires only the availability of genomic sequence and exon predictions, from which probes can be rapidly and efficiently synthesized onto an array. The flexibility and short time scale for designing and fabricating exon and tiling arrays using the ink-jet platform could substantially accelerate gene discovery.

Finally, our approach could be useful in the identification and analysis of genes underlying complex diseases. Genetic linkage studies of polygenic traits typically yield a dozen loci, each up to 20–30 megabases long. It is feasible to construct tiling arrays across all loci and probe them with mRNA samples from relevant normal and diseased tissues to ascertain both gene content and activity. Such analyses may provide not only more direct routes to the culpable genes, but also have the potential to uncover regulatory mutations by observed alterations in gene activity. □

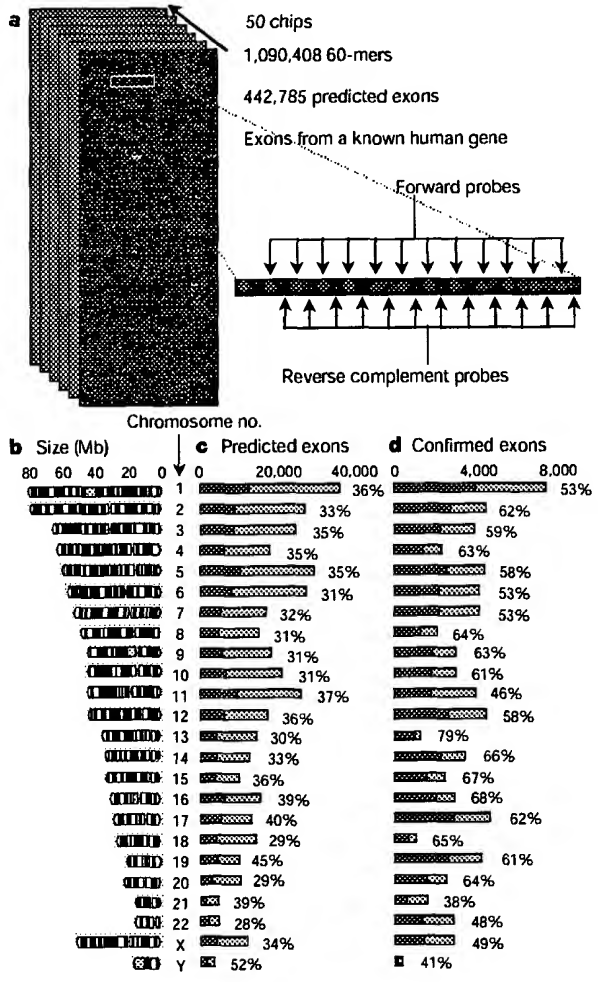

Figure 4 Whole-genome scan for validating predicted exons. **a**, Fifty 1 x 3-inch ink-jet arrays were used to test 442,785 Genscan predicted exons under two conditions. For each predicted exon, the best one or two 60-mer probes were selected, resulting in the set of 1,090,408 probes which were distributed over 50 arrays (~25,000 60-mers per array). The arrays also included 110,000 reverse complement probes and 48,500 control probes. The arrays were hybridized with Cy-3 or Cy-5 labelled mRNA from two human cell lines (Raji and Colo). Enlarged image, probes representing exons from a known gene with alternating forward and reverse complement probes. All experiments were performed in duplicate with a fluor reversal (100 arrays total). **b**, The sizes of the 24 human chromosomes (left). **c**, The number of predicted exons that were experimentally verified (red bars) for each of the chromosomes. Grey bars, number of predicted exons on each chromosome. **d**, A similar analysis for the confirmed exons across the human genome.

Methods

Sources of predicted exons

To analyse chromosome 22q, we designed a single ink-jet oligonucleotide microarray to represent 8,183 sequences that had been identified or confirmed as having coding potential (Sanger Centre). We used two sources of information: 6,650 Genscan-predicted exon sequences, and 3,381 validated exon sequences identified by aligning the first complete version of the human chromosome 22 sequence with sequences from experimentally validated transcripts[2]. Of this set of 10,031 exons, 1,847 had coordinates identical to those of other exons and were removed from the sequence pool. The remaining 8,183 exon sequences were subjected to an oligonucleotide design process to identify the two best candidate probes for a given exon sequence (see below). For the whole-genome exon scan, we designed ink-jet oligonucleotide microarrays to 442,785 predicted exons selected from the publicly available assembled sequence in the Ensembl database as of 15 June 2000. Specifically, we selected 554,202 non-redundant sequences from an initial set of 628,635 Genscan predicted exons[24]. We removed 111,417 more sequences from the list after they were flagged by the RepeatMasker algorithm (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker).

Probe selection for the exon-scanning arrays

For each of the predicted exons, we selected the top two 60-mers using an algorithm that takes into account binding energies, base composition, sequence complexity, cross-hybridization binding energies and secondary structure. For exon sequences of 60 nucleotides or less, we designed a single probe consisting of the entire exon sequence. For the 8,183 predicted exons on chromosome 22, 15,511 60-mers were selected and printed on a single array. For the whole-genome exon arrays, we selected 1,090,408 60mers to represent the 442,785 GenScan predicted exons from the Ensembl database. For 78,486 of the exons annotated as 'confirmed', the reverse-complement probes were also selected and placed next to the regular probes on the array as negative controls.

Probe selection for tiling arrays

In the tiling experiment described in Fig. 3, 60-mer probes were placed in 10-bp intervals across a 113.8-kb region of chromosome 22 containing the novel testis transcript described in Fig. 2e (BAC clone AL031587). The reverse complements for each of the tiling probes were also included on the array to allow transcripts on either strand to be detected. The genomic sequences used in the tiling experiments were repeat-masked before probe selection but no other exclusionary filters were applied.

Array synthesis

We synthesized the oligonucleotide arrays on 1 × 3-inch glass slides using ink-jet technology[26]. The phosphoramidite monomers were delivered by a standard ink-jet printer head to defined positions on a glass surface containing exposed hydroxyl groups. The remaining synthesis steps are similar to traditional oligonucleotide synthesis. Using this approach, up to 25,000 different 60-mers can be synthesized on a single slide. Around 1,000 'gridline' probes (5′ CCTATGTGACTGGTCGATGCTACTA 3′) are placed around the perimeter of each array. Fluorescently labelled synthetic oligonucleotides complementary to the control probes are included in all hybridizations. Arrays based on Rosetta designs were purchased from Agilent Technologies.

Preparation of labelled cDNA

We used the following human cell lines: Jurkat (T lymphocyte, ATCC no. TIB-152), K562 (chronic myelogenous leukaemia, ATCC no. CCL-243), Raji (Burkitt's lymphoma, ATCC no. CCL-86), Colo (colorectal adenocarcinoma, ATCC no. CCL-220), 293 (embryonic kidney, ATCC no. CRL-1573.1) and HepG2 (hepatocellular carcinoma, ATCC no. CRL-11997). Poly-A^+ RNA (mRNA) was isolated from each of the cytoplasmic RNA samples as described[27]. The 'pool' RNA sample described in Fig. 2 contains an equal mixture of four human cell lines (Jurkat, K562, Raji and Colo). The 41 mRNA samples from the human tissues described in Fig. 2 were purchased from commercial sources and are described at www.rii.com/Publications. For a single hybridization, we combined 1.5 μg of mRNA with 1.0 μg of random 9-mers and incubated the mixture for 10 min at 70 °C, 5 min at 4 °C and 10 min at 22 °C. To this mixture we added 0.5 mM amino-allyl dUTP (Sigma A-0410), 0.5 mM dNTP, 1 × RT buffer, 5 mM $MgCl_2$, 10 mM DTT and 200 units of Superscript (GibcoBRL), bringing the final reverse transcription reaction volume to 40 μl. This reverse transcription reaction was incubated for 20 min at 42 °C and the RNA was hydrolysed by adding 20 μl EDTA + NaOH and incubating at 65 °C for 20 min. The reaction was

neutralized by adding 20 µl of 1M Tris-HCl pH 7.6. We concentrated the resulting amino-allyl labelled single-stranded cDNA using a Microcon-30 (Millipore), and coupled it to Cy3 or Cy5 dye using a CyDye kit (Amersham Pharmacia Q15108). The per cent dye incorporation and total cDNA yield were determined spectrophotometrically. Pairs of Cy5/Cy3-labelled cDNA samples were combined and hybridized as described[22].

Analysis and visual display of expression data

Array images were processed as described[22] to obtain background noise, single channel intensity and associated measurement error estimates. Expression changes between two samples were quantified as $\log_{10}$ (expression ratio) where the 'expression ratio' was taken to be the ratio between normalized, background-corrected intensity values for the two channels (red and green) for each spot on the array. An error model for the log ratio was applied[22] to quantify the significance of expression changes between two samples. The colour displays in Fig. 2 show $\log_{10}$(expression ratio) as red when the red channel is upregulated relative to the green channel, green when the red channel is downregulated relative to the green channel, black when $\log_{10}$(expression ratio) is close to zero, and grey when data from one or both of the channels for a given probe are unreliable.

Identifying EVGs by co-regulation

Exons were grouped into EVGs by a two-step gene identification algorithm. First, each probe was assigned a similarity measure, based on the moving average (using a window size equal to six probes) of pair-wise Pearson correlation coefficients between the log ratios of probe intensities in neighbouring exons. Probes with correlation coefficients above 0.5 in a given window were selected as seeds for EVGs. The 0.5 threshold and window size were determined empirically by training the model on a subset of the known chromosome 22 genes. Second, probes neighbouring a seed region were merged into the region if the pair-wise correlation coefficients between the neighbouring probe and the average in the seed region exceeded 0.5. This process continued, allowing for gaps between probe pairs to account for failed probes and/or false exon predictions (gaps were not allowed to exceed five probes), until no probes flanking the candidate region met the significance threshold of correlation with the exon cluster. The final exon clusters resulting from the gene detection algorithm were identified as an EVG. Not all condition pairs (rows) were considered in forming EVGs. Elements in a given row had to have significant P values (≤ 0.01) to be included in the analysis. Once an EVG was formed, the colour display (as in Fig. 2) was updated by reordering the condition pairs according to a hierarchical clustering algorithm, as described[28].

Annotation of EVGs

Predicted transcripts for all EVGs identified from the chromosome 22 exon data across the 69 condition pairs were formed by combining the individual exons into a single sequence. Each of these sequences was searched against dbEST and the NR protein databases using gapped BLASTN and BLASTX (www.ncbi.nlm.nih.gov), respectively, to determine the extent to which the EVG sequences were similar to other sequence data. We declared sequences similar if the corresponding E-value for the alignment was less than 10^{-9}, using default parameters for gapped BLAST. BLAST results were used to determine the degree of sequence support defining a predicted transcript. These results were also used to determine the degree of existing sequence support for each of the EVGs detected from the chromosome 22 exon arrays.

Quantitative analysis of whole-genome exon data

We used an intensity-based algorithm to verify predicted exons experimentally across the entire human genome. Specifically, we used raw intensity measurements for the forward-strand (FS) probes and the corresponding raw intensity measurements for the reverse-complement (RC) probes in conjunction with the respective standard deviations of those measurements to determine the significance of the FS probe intensities. We controlled for nonspecific cross-hybridization using RC probes, given that the reverse complement of a DNA sequence has equivalent sequence complexity to the forward strand sequence with respect to a variety of measures (such as GC content and GC trend). An exon was called 'present' if the intensity difference between an FS probe and the RC probe had $P < 0.01$ in either the red or green channel, and if the FS probe intensity had a $P < 0.01$ for being above background in the channel where the difference was considered most significant. P values were calculated using a t-test applied to the difference of the mean pixel intensities and to the difference of the mean FS/background intensities.

These single channel exon detection methods were applied only to those exons in which reverse-complement probes were designed. In the remaining cases, the significance of the single channel intensities was determined using the above-background criterion described above. We applied a correction to the detection percentages given for the predicted exons listed in Fig. 4, based on false positive estimates for above-background calls that were determined using the FS/RC probe intensity difference calls for the confirmed exons. Error models used in this analysis to assess ratio significance were as described[28]. Of the 88,374 confirmed exons represented on the genome-wide exon arrays, 78,486 had corresponding RC probes. To assess the rate of false positives expected in the single-channel assessments, we used a similar detection procedure to determine the number of RC probe intensity measurements that were significantly greater than the corresponding FS probe intensity. Our results indicate that the false positive rate of detection using the single channel method was ~ 5%.

Received 28 November 2000; accepted 9 January 2001.

1. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
3. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
4. Wheelan, S. J. & Boguski, M. S. Late-night thoughts on the sequence annotation problem. *Genome Res.* **8**, 168–169 (1998).
5. Boguski, M. S. Biosequence exegesis. *Science* **286**, 453–455 (1999).
6. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
7. Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735–1744 (1997).
8. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
9. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
10. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
11. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
12. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
13. Marshall, E. Public-private project to deliver mouse genome in 6 months. *Science* **290**, 242–243 (2000).
14. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
15. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNAs collection. *Nature* **409**, 685–690 (2001).
16. Hanke, J. *et al.* Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* **15**, 389–390 (1999).
17. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293 (1999).
18. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367–370 (2000).
19. Brett, D. *et al.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474**, 83–86 (2000).
20. de Souza, S. J. *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA* **97**, 12690–12693 (2000).
21. Penn, S. G., Rank, D. R., Hanzel, D. K. & Barker, D. L. Mining the human genome using microarrays of open reading frames. *Nature Genet.* **26**, 315–318 (2000).
22. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
23. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
24. Hubbard, T. & Birney, E. Open annotation offers a democratic solution to genome sequencing. *Nature* **403**, 825 (2000).
25. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
26. Blanchard, A. P., Kaiser, R. J. & Hood, L. E. High-density oligonucleotide arrays. *Biosens. Bioelectron.* **6/7**, 687–690 (1996).
27. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).
28. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).

Supplementary Information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank S. H. Friend for encouragement and support and J. Rine, M. V. Olson, C. Roberts and T. Hughes for critical readings of the manuscript.

Correspondence and requests for materials should be addressed to M.S.B. (e-mail: msb@rii.com).

Mining the human genome using microarrays of open reading frames

Sharron G. Penn[1]*, David R. Rank[1]*, David K. Hanzel[1] & David L. Barker[2]

*These authors contributed equally to this work.

To test the hypothesis that the human genome project will uncover many genes not previously discovered by sequencing of expressed sequence tags (ESTs), we designed and produced a set of microarrays using probes based on open reading frames (ORFs) in 350 Mb of finished and draft human sequence. Our approach aims to identify all genes directly from genomic sequence by querying gene expression. We analysed genomic sequence with a suite of ORF prediction programs, selected approximately one ORF per gene, amplified the ORFs from genomic DNA and arrayed the amplicons onto treated glass slides. Of the first 10,000 arrayed ORFs, 31% are completely novel and 29% are similar, but not identical, to sequences in public databases. Approximately one-half of these are expressed in the tissues we queried by microarray. Subsequent verification by other techniques confirmed expression of several of the novel genes. Expressed sequence tags (ESTs) have yielded vast amounts of data[1,2], but our results indicate that many genes in the human genome will only be found by genomic sequencing.

We downloaded for analysis all human genomic clones greater than 50 kB in size, spanning less than 10 contigs, and submitted to GenBank between 15 May and 15 October 1999. This corresponds to 2,354 clones, or approximately 350 Mb. After masking repetitive elements, we sought ORFs using three gene-finding algorithms developed on independent training sets: Grail (which uses a neural network; ref. 3), Genefinder (which uses a hidden Markoff model; ref. 4) and DiCTion (which searches for coding regions based on Fourier transform methods; J. Graham, pers. comm.). Gene-finding programs are notorious for their generation of false-positive results[5]. The question thus arises: are these predicted coding regions real? To minimize this problem we required at least two independent predictions[6] of a coding region as a criterion for evaluation by microarray analysis. We collected the consensus ORFs into putative gene 'bins' using an empirical criterion. We then chose the largest ORF from each bin that did not contain any repetitive sequence. We also selected all consensus ORFs greater than 500 bp. We thereby attempted to approximate one exon per gene, but a number of genes were represented by multiple elements (Fig. 3; and Fig. A, see http://genetics.nature.com/supplementary_info/), and it is highly probable that we have missed some genes. Primers were designed to PCR-amplify 500-bp sequences centred around the ORF of interest, and universal primer sequences were added to the 5′ end of each ORF-specific primer.

We observed a mean exon size of 229 bp and a median size of 150 bp (n=9498). In contrast, the amplicons were designed to have a very narrow distribution (475±25 bp), facilitating uniform retention on the microarray[7]. Thus, approximately 50% of the average PCR product represents a coding region. We found that larger exons tended to yield absolute hybridization signals of greater magnitude (data not shown), and, for this reason, sought to avoid using short exons as probes. The ORFs were PCR amplified from genomic DNA and their sizes confirmed by agarose-gel electrophoresis. We successfully sequenced 65% of all PCR amplicons, confirming their identity. Most 'unsuccessful' sequences were due to failed PCR reactions; others yielded poor sequence data. The reasons for this are unclear, but may be related to the quality of early draft sequence including misassemblies, inclusion of vector and host contamination. We thus distilled 350 Mb of genomic DNA to 9,498 amplicons, and spotted them in duplicate onto treated glass slides.

All arrayed gene element sequences were subjected to BLAST (ref. 8) analysis against the GenBank databases (7 May 1999; release 2.0.9): 40% of the sequences obtained an exact match (E values<1 e^{-100}) with either an EST or a known mRNA; 29% showed some homology with a known EST or mRNA (E values=1 e^{-5} to 1 e^{-99}). The remaining 31% of the elements showed no sequence homology with GenBank sequences. It should be noted, however, that as our selection process does not favour central or terminal exons, it is likely that some of the apparently novel exons are parts of genes whose 3′ or 5′ ends have been captured as ESTs.

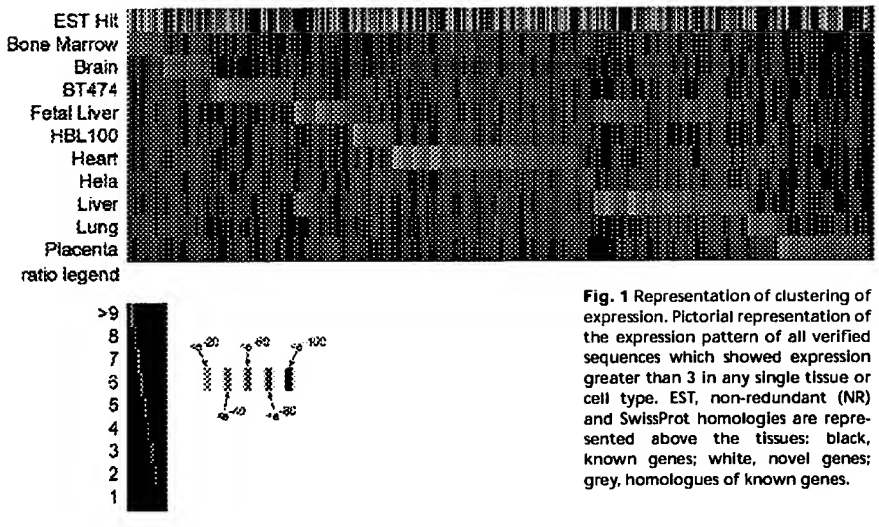

Fig. 1 Representation of clustering of expression. Pictorial representation of the expression pattern of all verified sequences which showed expression greater than 3 in any single tissue or cell type. EST, non-redundant (NR) and SwissProt homologies are represented above the tissues: black, known genes; white, novel genes; grey, homologues of known genes.

[1]*Advanced Research Team, Molecular Dynamics Inc., Sunnyvale, California, USA.* [2]*Illumina, Inc., San Diego, California, USA. Correspondence should be addressed to D.R.R. (e-mail: David.Rank@am.apbiotech.com).*

Table 1 • Brain-specific transcripts identified by microarray

Chip sequence	Homology to known EST	Protein function as ascribed by GenBank
AP000217-1	high	S-100 protein, β-chain, Ca^{2+} binding protein expressed in central nervous system[11]
AP000047-1	high	unknown function
AC006548-9	high	similar to mouse membrane glycoprotein M6, expressed in central nervous system
AC007245-5	high	similar to amphiphysin, a synaptic vesicle-associated protein[12]
L44140-4	high	endothelial actin-binding protein found in non-muscle filamin
AC004689-9	high	protein phosphatase PP2A; neuronal; downregulates activated protein kinases[13]
AL031657-1	high	unknown function; contains ankyrin motif
AC009266-2	low	low homology to the synaptotagmin I protein in rat[14]
AP000086-1	low	unknown, very poor homology with collagen
AC004689-3	high	Protein phosphatase PP2A, neuronal; downregulates activated protein kinases[13]

The microarrays were hybridized with RNA samples obtained from seven tissues and three cell lines. Each mRNA was reverse-transcribed to Cy3-cDNA. A pool of the RNAs obtained from all ten sources (that is, the tissues and cell lines) was transcribed to Cy5-cDNA and used as a reference target. Whereas this strategy allowed us to survey a large number of tissues, it attenuated the measurement of relative gene expression, as every highly expressed gene in the tissue or cell type channel will be present at a level of at least 10% in the control channel. For this reason, we represented data both in terms of normalized ratios and normalized signal intensity. Of 9,498 arrayed elements on 2 chips (including positive and negative controls and 'failed' products), 4,800 (51%) were expressed in at least 1 tissue or cell type. Of the gene elements showing significant signal, approximately 1,870 (39%) were expressed in all 10 tissue or cell types. In the next most common grouping, approximately 720 (15%) were expressed in only a single tissue or cell type.

We further analysed the genes expressed predominantly in a single tissue (Fig. 1) and found that a large fraction of them are novel, even in tissues that have been exhaustively studied by EST sequencing.

We used several approaches to validate our methods. We included BAC AC006064 in the sequence data set, as it is known to contain the gene encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH). The algorithms selected 25 sequences from BAC AC006064 for spotting onto the array, of which 4 corresponded to *GAPDH*. A commercially available *GAPDH* cDNA was also spotted onto the array: hybridization with labelled target showed excellent agreement between expression detected by probes representing single exons (with an average length of 229 bp) and that detected by the full-length cDNA control (of 1,100 bp; data not shown).

To confirm the novel gene sequences, we selectively assayed gene expression by PCR from a panel of commercially available cDNAs derived from a variety of tissue mRNAs. The primers used were selected from those used to generate the microarray probes. We selected the elements according to the following criteria: (i) they were previously absent from public databases as coding sequences; (ii) they sequenced successfully; and (iii) they yielded interesting tissue-specific gene-expression patterns as measured by microarray. We thereby confirmed that AL079300-1 is expressed in cardiac tissue, and AL031734-1, in placental tissue. Neither was found to be expressed by other tissues analysed by microarray.

Clearly, all microarray results cannot be confirmed by independent methodologies. Evidence supporting our data, however, exists in public databases. For example, we carried out an analysis of the amplicons that showed high signal only in brain, of which there were 82. Of the ten with the highest degree of expression, six are known to have a role in the central nervous system or brain (Table 1). We then sought matches for the sequences generating the greatest (normalized) signal intensities in brain, regardless of expression in other tissues. Of the 20 with highest expression in the brain, 3 were similar to the gene encoding tubulin (AC008079-5, AF146191-2, AC007664-4), 2 were similar to the gene encoding actin (AL035701-2, AL034402-1) and 5 were homologous with *GAPD* (AL035604-1, Z86090-1, AC006064-L, AC006064-K, AL035604-3). Expression patterns across multiple tissues were also confirmed. For example, sequence L29074-7, which encodes a ferritin heavy chain protein, is reported to be expressed in brain and liver[9], consistent with our results obtained by microarray.

On completing sequence analysis of chromosome 22, Dunham *et al.*[10] predicted that it contains at least 679 genes. Our study analysed 49% (16 Mb) of the chromosome-22 sequence. Assuming we predicted 1 exon per gene, we found 298 genes that were also predicted by Dunham *et al.*, which implies that we would have identified approximately 90% of the genes identified by Dunham *et al.* had we analysed the entire chromosome. We also found 235 additional potential genes not included in the minimal set of Dunham *et al.* Whereas some of these will be duplicates (generated by two ORFs from distant parts of the same gene) and false positives, many will be real (see example AL079300-1 above).

For each genomic clone analysed by microarray, we accumulate a plethora of information. We have therefore devised a visualization tool to present this information, which we call a 'Mondrian' (in deference to the Belgian painter; Fig. 2). A 'Mondrian' of a BAC encompassing the gene encoding carbamyl phosphate synthetase illustrates the high degree of confidence with which single exons may be used to query gene expression (Fig. A, see http://genetics.nature.com/supplementary_info/), whereas a Mondrian of another BAC (Fig. 3) illustrates the power of the strategy for detecting genes.

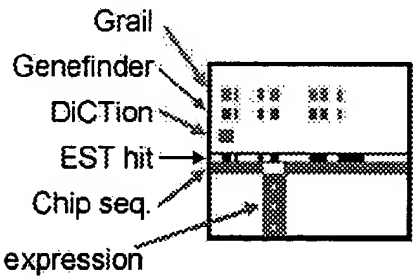

Fig. 2 A 'Mondrian' of a 'virtual' BAC. The red line running left to right depicts clone sequence. Results from exon-finding programs are represented above the red line (blue, GeneFinder; green, Grail II; grey, DiCTion). Sequences were selected as described in the text, and depicted as a white bar masking out the red line. Information about homology of the sequence within GenBank is depicted above the red line, when available. Black indicates 'known' and white indicates unknown regions. Microarray expression data is represented below the red line. A colour depiction of ratio-based gene expression for expression in three tissues is shown in this region. Shades of green depict elevated expression in that sample over the control and shades of red depict higher signal in the control sample. Darker shades of either red or green indicate higher or lower ratios of gene expression compared with the control. Within each bar, a white circle is drawn, its size proportional to the size of the signal intensity.

The preparation of DNA microarrays by spotting gene fragments directly amplified from genomic DNA is a new method of gene discovery, and contrasts with the standard practice of using cDNA libraries for generating microarray probes in that the amplified fragments are thoroughly characterized and of uniform length, and previously unknown genes can be analysed for differential gene expression. We note, however, that measuring expression of each exon in a gene may also be an effective method for detecting alternate splicing and determining its tissue-specific pattern. Application of the technique allows discovery and characterization of genes in new genomic sequence in advance of isolation and cloning. Data are available at http://www.hmorf.com.

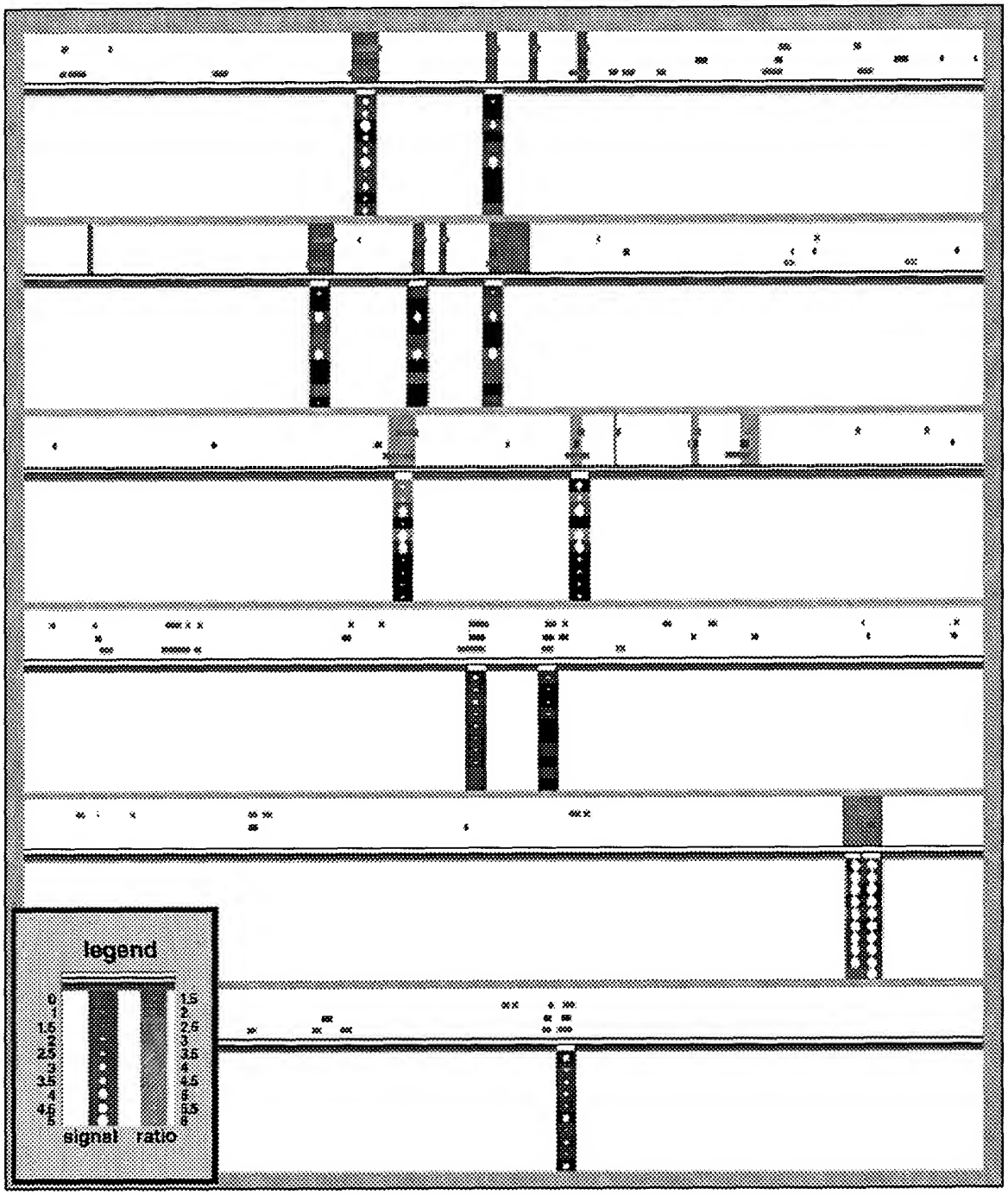

Fig. 3 A study of BAC AL049839, depicted by 'Mondrian'. We selected 12 exons from this BAC, of which 10 were successfully amplified and sequenced. The ten exons we arrayed represent four to six genes (four known and two putative genes). The four known genes are protease inhibitors. These data show that exons selected from the same gene display the same expression patterns, depicted below the red line. A novel gene is also found from 86.6 kb to 88.6 kb, in agreement with all the exon-finding programs. Similar expression patterns and their proximity to one another suggests the two exons are common to a single gene. Red represents the kallistatin protease inhibitor (P29622); purple represents the serine protease inhibitor (P05154); turquoise represents α1 anti-chymotrypsin (P01011); and mauve represents 40S ribosomal protein (P08865). Each panel represents 25 kb.

Methods

Preparation of labelled cDNA. Human mRNA samples were purchased from commercial sources (Clontech and ATCC): heart, brain, liver, fetal liver, placenta, lung, bone marrow, and the cell lines HeLa S3, BT 474 (human breast ductal carcinoma cell line) and HBL 100 (human breast cell line). Cy3-dCTP and Cy5-dCTP (Amersham Pharmacia Biotech (APB)) were incorporated into cDNA during reverse transcription as follows. Poly(A) mRNA (1 μg), Oligo(dT)$_{12-18}$ primer (1 μg) and random 9-mer primers (2 μg) in a volume of 11 μl were incubated at 70 °C for 10 min. After snap cooling on ice, the following was added to the RNA to the stated final concentration and to a volume of 20 μl: 1×Superscript II buffer, 0.01 M DTT, 100 μM dATP, 100 μM dGTP, 100 μM dTTP, 50 μM dCTP, 50 μM Cy3-dCTP or Cy5-dCTP, and 200 U Superscript II enzyme. The reaction was incubated for 2 h at 42 °C. After incubation, the cDNA was isolated by adding 1 U Ribonuclease H and incubating for 30 min at 37 °C. The reaction was then purified using a Qiagen PCR cleanup column. Probe was eluted using Tris HCl (10 mM, pH 8.5).

Hybridization. Dye incorporation and total cDNA were determined spectrophotometrically. A volume of probe equivalent to 50 pmoles of Cy3 and Cy5 dye was dried and resuspended in 30 μl hybridization solution (50% formamide, 5×SSC, 0.2 μg/μl poly(dA), 0.2 μg/μl human Cot1 DNA, 0.5% SDS). Hybridizations were carried out under a coverslip, and the array placed in a humid oven at 42 °C overnight. Slides were washed in 1×SSC, 0.2% SDS at 55 °C for 5 min, followed by 0.1×SSC, 0.2% SDS, at 55 °C for 20 min, then were briefly dipped in water and dried thoroughly under a gentle stream of nitrogen. Slides were scanned on a Molecular Dynamics Generation III scanner[7].

Normalization of microarray data. By 'balancing' dye load before hybridization we limit the amount of normalization necessary, because signal intensities in each channel become equivalent. Molecular Dynamics spotting instrumentation spots each sample in duplicate[7]. Ratios were normalized as follows: the average ratio of the two hybridization signals from the duplicate spots is accepted for further analysis only if the duplicate ratios are within 25%. The average ratios of all the accepted data on each slide falls within a normalized distribution. The ratios are normalized by dividing all ratios by the average ratio of all the spots on the slide. This normalizes the average ratio around 1, and allows ratios to be compared between slides. Signal is normalized by dividing all signals by the average signal of the spots on the slide. This results in the 'average' spot having a normalized signal of 1. We define significant expression as signal three times greater than biological noise (average signal from the negative control, *Escherichia coli* genes).

Microarray controls. Positive and negative controls were spotted on the microarray slides. The positive controls were probes encoding GAPDH and actin, and Human Cot-1 DNA. The negative controls were salmon sperm DNA and *Escherichia coli* DNA.

Sequencing. All PCR products were sequenced using energy transfer dye terminators (APB) on the Molecular Dynamics MegaBACE using standard protocols.

Acknowledgements

We thank D. Jenkins and other members of the ART team for assistance, support and encouragement; R. Thomas for programming assistance; Operon for cooperation on this project; and J. Graham for the use of DiCTion.

Received 14 March; accepted 18 August 2000.

1. Strausberg, R.L., Dahl, C.A. & Klausner, R.D. New opportunities for uncovering the molecular basis of cancer. *Nature Genet.* **15**, 415–416 (1997).
2. Adams, M.D. *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3–174 (1995).
3. Uberbacher, E.C. & Mural, R.J. Locating protein-coding regions in human DNA sequnce by multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88**, 11261–11265 (1991).
4. Solovyev, V.V., Salamov, A.A. & Lawrence, C.B. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22**, 5156–5163 (1994).
5. Burset, M. & Guigó, R. Evaluation of gene structure prediction programs. *Genomics* **34**, 353–367 (1996).
6. Ansari-Lari, M.A. *et al.* Comparative sequence analysis of the gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res.* **8**, 29–40 (1998).
7. Worley, J. *et al.* A systems approach to fabricating and analyzing DNA microarrays. in *Microarray Biochip Technnology* (ed. Schena, M.) 65–86 (Biotechniques Books, Natick, Massachusetts, 2000).
8. Altschul, S.F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
9. Joshi, J.G., Fleming, J.T., Dhar, M. & Chauthaiwale, V. A novel ferritin heavy chain messenger ribonucleic acid in human brain. *J. Neurol. Sci.* **134**, 52–56 (1995).
10. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
11. Heizmann, C.W. Ca2+-binding S100 proteins in the central nervous system. *Neurochem. Res.* **9**, 1097–2000 (1999).
12. Wigge, P. & McMahon, H.T. The amphiphysin family of proteins and their role in endocytosis at the synapse. *Trends Neurosci.* **21**, 339–344 (1998).
13. Millward, T.A., Zolnierowicz, S. & Hemmings, B.A. Regulation of protein kinase cascades by protein phosphatase 2A. *Trends Biochem. Sci.* **24**, 186–191 (1999).
14. Ullrich, B. *et al.* Functional properties of multiple synaptotagmins in brain. *Neuron* **6**, 1281–1291 (1994).

Atty. Docket No.: AEOMICA-1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: PENN, Sharron G. *et al.*

Serial No.: 09/774,203 Confirm. No.: 7320

Filed: January 29, 2001

For: METHODS AND APPARATUS FOR PREDICTING, CONFIRMING, AND DISPLAYING FUNCTIONAL INFORMATION DERIVED FROM GENOMIC SEQUENCE

Sunnyvale, CA

Hon. Commissioner for Patents
Washington, D.C. 20231

DECLARATION OF SHARRON G. PENN, Ph.D.
IN SUPPORT OF PETITION TO MAKE SPECIAL

Sir:

I, SHARRON G. PENN, declare and state as follows:

1. I am an inventor of the inventions described and claimed in United States Patent application serial no. 09/774,203 ("the '203 application"), and the first-named author of Penn *et al.*, "Mining the human genome using microarrays of open reading frames," *Nature Genetics* 26:315 - 318 (November 2000), a publication that includes many of the data set forth in the '203 application. I am the director of research and development and director of production at Aeomica, Inc., the assignee of the above-identified patent application and of the inventions

described and claimed therein. I believe that I have a good knowledge of the pertinent prior art. My *curriculum vitae* is attached hereto as Exhibit 1.

2. I have read Shoemaker *et al.*, "Experimental annotation of the human genome using microarray technology," *Nature* 409:922 - 927 (2001), which cites our *Nature Genetics* publication. I believe that Shoemaker *et al.* and Penn *et al.* are the two references most closely related to the subject matter encompassed by the claims of the '203 application.

3. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of the application or any patent that issues thereon.

8— 14— 01

Date

Sharron G. Penn, Ph.D.

Sharron Gaynor Penn, Ph.D.

Aeomica Inc.
Mailstop #9
928 East Arques Avenue
Sunnyvale, CA 94086
Tel: (408) 737 3060
Email: Sharron.Penn@am.apbiotech.com

EDUCATION

1995- 1997	University of California, Davis Post- doctoral fellow in the laboratories of Prof. Carlito Lebrilla.
1991-1994	University of York, England. Ph.D. entitled " Chiral Analysis by Capillary Electrophoresis" under the supervision of Dr. David M. Goodall. Funded by Pfizer Central Research through a CASE award.
1987-1991	Nottingham Trent University (formerly Trent Polytechnic), England. B.Sc (Hons) Applied Chemistry, class II(I). Third year spent at an industrial placement with Merck Sharp and Dohme.

AWARDS

April 2001	Awarded the Ellen Weaver Award by Northern California Chapter of Association of Women in Science.
December 1994	Awarded the Ronald Belcher Memorial Lectureship by The Royal Society of Chemistry.
September 1994	Awarded the PASG 1994 award by the UK Pharmaceutical and Analytical Sciences Group.
January 1994	Awarded the Kathleen Mary Stott prize in Chemistry for Graduate Research by the University of York.

WORK EXPERIENCE

April 2000-Present	Director, Research and Development, and Director of Production, Aeomica, Inc.
April 1996-2000	Scientist, Advanced Research Team, Molecular Dynamics, Sunnyvale, CA.
August 1994	Course tutor, " Hyphenated Capillary Electrophoresis-Mass Spectrometry techniques" , University of York Capillary Electrophoresis short course.
Sept 1989-Sept 1990	Analytical Chemist, Medicinal Chemistry Department, Merck Sharp and Dohme Neuroscience Research Center, Harlow, England.

PUBLICATIONS

25
Bartosiewicz MJ, Jenkins D, Penn S, Emery J, Buckpitt A. (2001)
"Unique gene expression patterns in liver and kidney associated with exposure to chemical toxicants."
J Pharmacol Exp Ther, 297,895-905

24
Bartosiewicz M., Penn S., Buckpitt A., (2001)
"Application of Gene Arrays in environmental toxicology : fingerprints of gene regulation associated with cadmium chloride, benzo(a)pyrene, and trichloroethylene"
Environmental Health Perspectives 109, 71-72.

23
J. Worley, K. Bechtol, S. Penn, D. Roach, D. Hanzel, M. Trounstine and D. Barker. (2000)
A Systems Approach to Fabricating and Analyzing DNA Microarrays, p. 65-86. Chapter 4 in Microarray Biochip Technology, Edited by Mark Schena.

22
Penn S.G., Rank D.R., Hanzel D.K., Barker D.L. (2000)
"Mining the human genome using microarrays of open reading frames."
Nature Genetics, 26, 315-318.

21
Penn, S.G., Costa, D.A., Balch, A.L., Lebrilla, C.B. (1997)
" Analysis of C-60 oxides and C120On (n=1,2,3) using matrix assisted laser desorption-ionization Fourier transform mass spectrometry" . International Journal of Mass Spectrometry and Ion Processes, 169, 371-386.

20
Cancilla, M.T., Penn, S.G., Lebrilla, C.B. (1998)
" Alkaline degradation of oligosaccharides coupled with matrix-assisted laser desorption/ionization Fourier transform mass spectrometry: A method for sequencing oligosaccharides" . Analytical Chemistry, 70, 663-672.

19
Wisner, E.R., AhoSharon, K.L., Bennett, M.J., Penn, S.G., Lebrilla C.B., Nantz M.H., (1997)
" A modular lymphographic magnetic resonance imaging contrast agent: Contrast enhancement with DNA transfection potential." Journal of Medicinal Chemistry 40, 3992-3996.

18
Penn, S.G., Hu, H.N., Brown, P.H., Lebrilla, C.B. (1997)
" Direct analysis of sugar alcohol borate complexes in plant extracts by matrix-assisted laser desorption/ionization Fourier transform mass spectrometry" Analytical Chemistry, 69, 2471-2477.

17
Tseng K., Lindsay L.L., Penn S.G., Hedrick J.L., Lebrilla C.B. (1997)
" Characterization of neutral oligosaccharide-alditols from Xenopus laevis egg jelly coats by matrix-assisted laser desorption Fourier transform mass spectrometry." Analytical Biochemistry, 250, 18-28.

16
Penn, S.G., Cancilla, M.T., Green, M.K., Lebrilla, C.B. (1997)
" Direct comparison of matrix-assisted laser desorption/ionisation and electrospray ionisation in the analysis of gangliosides by Fourier transform mass spectrometry. European Mass Spectrometry, 3, 67-79.

15
Penn, S.G., He, F., Green, M.K., Lebrilla, C.B. (1997)
" The use of heated capillary dissociation and collision-induced dissociation to determine the strength of noncovalent bonding interactions in gas-phase peptide-cyclodextrin complexes." Journal of the American Society for Mass Spectrometry, 3, 244-252.

14
Hu, H.N., Penn, S.G., Lebrilla, C.B., Brown, P.H. (1997).
" Isolation and characterization of soluble boron complexes in higher plants - The mechanism of phloem mobility of boron." Plant Physiology, 113, 649-655.

13
Gard, E.E., Green, M.K., Warren, H., Camara, E. J. O., Penn S.G., Lebrilla C.B. (1996)
" A dual vacuum chamber Fourier transform mass spectrometer with rapidly interchangeable FAB, MALDI and ESI sources: Electrospray results." International Journal of Mass Spectrometry and Ion Processes, 158, 115-127.

12
Franklin, M.A., Penn, S.G., Lebrilla, C.B., Lam, T.H., Molinski T.F. (1996)
" Bastadin 20 and bastadin O-sulfate esters from Ianthella basta: Novel modulators of the Ry(1)R FKBP12 receptor complex." Journal of Natural Products, 59, 1121-1127.

11
Camara E., Green M.K., Penn S.G., Lebrilla C.B., (1996)
" Chiral Recognition is observed in the deprotonation reaction of Cytochrome C by (2R)- and (2S)-2-butylamine" . Journal of the American Chemical Society, 118, 8751-8752.

10
Penn S.G., Cancilla M.T., Lebrilla C.B., (1996)
" Collision-induced dissociation of branched oligosaccharide ions with analysis and calculation of relative dissociation thresholds" Analytical Chemistry, 68, 2331-2339.

9
Cancilla M.T., Penn S.G., Carroll J.A., Lebrilla C.B. (1996)
" Coordination of alkali metals to oligosaccharides dictates fragmentation behavior in matrix assisted laser desorption ionization mass spectrometry." Journal of the American Chemical Society, 118, 6736-6745.

8
Carroll J.A., Penn S.G., Fannin S.T., Wu J.Y., Lebrilla C.B. (1996)
A dual vacuum chamber Fourier transform mass spectrometer with rapidly interchangeable LSIMS, MALDI and ESI sources – initial results with LSIMS and MALDI" Analytical Chemistry, 68, 1798-1804.

7
Green M.K., Penn S.G., Lebrilla C.B. (1995)
" The complexation of protonated peptides with saccharides in the gas phase decrease the rates of hydrogen/deuterium exchange reactions" . Journal of the American Society for Mass Spectrometry, 6, 1247-1251.

6
Piperaki S., Penn S.G., Goodall D.M., (1995)
" Systematic approach to treatment of enantiomeric separations in capillary electrophoresis and liquid chromatrography. 2. A study of the enatiomeric separation of fluoxetine and norfluoxetine." Journal of Chromatography A, 700, 59-67.

5
Penn S.G., Bergstrom E.T., Knights I., Liu G.Y., Goodall D.M. (1995)
" Capillary electrophoresis as a method for determining binding constants – application to the binding of cyclodextrins and nitrophenolates" Journal of Physical Chemistry, 99, 3875-3880.

4
Penn S.G., Liu G.Y., Bergstrom E.T., Goodall D.M., (1994).
" Systematic approach to treatment of enantiomeric separations in capillary electrophoresis and liquid chromatrography. 1. Initial evaluation using propranolol and dansylated amino acids" Journal of Chromatography A, 680, 147-155.

3
Penn S.G., Chiu R.W., Monnig C.A. (1994)
" Separation and analysis of cyclodextrins by capillary electrophoresis with dynamic fluorescence labeling and detection." Journal of Chromatography A, 680, 233-241.

2
Penn S.G., Bergstrom E.T., Goodall D.M., Loran J.S. (1994)
" Capillary electrophoresis with chiral selectors – Optimization of separation and determination of thermodynamics parameters for binding of tioconazole enantiomers to cyclodextrins" . Analytical Chemistry, 66, 2866-2873.

1
Penn S.G., Goodall D.M., Loran J.S. (1993)
" Differential binding of tioconazole enantiomers to hybroxypropyl-beta-cyclodextrin studied by capillary electrophoresis" . Journal of Chromatography A, 636, 149-152.